Variance Estimation for Fixed-Configuration, Systematic Sampling

Stephen V. Stehman
Biometrics Unit, Cornell University, Ithaca, NY 14850 USA

W. Scott Overton
Department of Statistics, Oregon State University, Corvallis, Oregon 97331 USA

BU-1010-M Revised August 1989

Variance Estimation for Fixed-Configuration, Systematic Sampling

Stephen V. Stehman and W. Scott Overton

Biometrics Unit, Cornell University,
and Department of Statistics, Oregon State University

Biometrics Unit Manuscript Bu-M 1010

Technical Report 134

August 1989

Abstract

Variable probability systematic sampling can be implemented on a universe in a particular order, called fixed-configuration sampling, or implemented on a randomly-ordered universe. The efficiency of the Horvitz-Thompson estimator for fixed-configuration sampling relative to random-order sampling depends on the relationship between the fixed-configuration and random-order pairwise inclusion probabilities, and the ordering of the ratios of the response variable to the auxiliary variable used to select the sample. The usual variance formulas for variable probability sampling are biased for estimating the fixed-configuration variance. To reduce the overestimation problem of the random-order variance estimators, an estimator analogous to the equal probability mean square successive difference estimator is proposed. The properties of this estimator and several other estimators of the fixed-configuraton variance are investigated both theoretically and empirically. While expected behaviors of the variance and variance estimators can be described, exceptional behaviors are to be expected because certain orderings of the universe will lead to anomalous results. These exceptions are difficult to predict because they are related to the pairwise inclusion probabilities.

Department of Statistics
Oregon State University
Corvallis, Oregon

1. INTRODUCTION

Systematic sampling is often advantageous over random sampling, whether in the framework of equal or unequal probability sampling. In addition to being easy to implement and often providing increased precision of estimators, systematic sampling furnishes information on spatial or temporal patterns in the population. Systematic sampling from an ordered universe is the focus of this paper. The variance and estimators of the variance of the Horvitz-Thompson estimator are investigated for the design of variable probability, systematic sampling.

The foundation of equal probability systematic sampling was given in the early papers of Madow and Madow (1944), Cochran (1946), Madow (1949), and Yates (1949). A review paper by Buckland (1951) summarized many of these early results. Quenouille (1949) and Das (1950) extended some theoretical results of one-dimensional systematic sampling to two-dimensions. Iachan (1982) and Bellhouse (1988) wrote review articles describing many of the results established after 1950.

Systematic sampling is widely accepted as a very practical sampling design, but unbiased estimation of variance is problematic. Two quotes, both made in reference to variable probability systematic sampling, sufficiently summarize a standard concern:

> "A drawback of the systematic method is, as usual, the absence of an unbiased estimate of the variance of the population total."
> Cochran (1977).

> "The main disadvantage of this method, [variable probability, systematic sampling] as in simple systematic sampling, is that it is not possible to get an unbiased variance estimator on the basis of a single sample."
> Murthy (1967).

Similar statements can be found in almost any sampling text, and in many quantitative methods books (cf. Freese (p. 61, 1962), Poole (p. 300, 1974)).

1.1 Notation

1. <u>General</u>

$\mathcal{A}$ = set of (i,j) pairs in the universe such that $\gamma_{ij} \neq 0$.

$\mathcal{A}'$ = the complement of $\mathcal{A}$, the set of (i,j) pairs in the universe such that $\gamma_{ij} = 0$.

δ^2 mean-square successive difference estimator

vps variable probability, systematic

2. <u>Summations</u>

$\sum$ summation over all units in the universe

$\sum_s$ summation over all units in the sample

$\sum\sum$ summation over all pairs of units in the universe

$\sum_s \sum_s$ ($j \neq i$) summation over all pairs of units in the universe excluding i=j

$\sum_{\mathcal{A}}$ represents summation over all pairs (i,j) in set $\mathcal{A}$ ($\mathcal{A}'$ can be substituted for $\mathcal{A}$ for summation over $\mathcal{A}'$).

$\sum_{\mathcal{A}}$ represents summation over all pairs (i,j) in the sample that are also in set $\mathcal{A}$ ($\mathcal{A}'$ can be substituted for $\mathcal{A}$ for summation over $\mathcal{A}'$).

3. <u>Observations</u>

y   response variable of interest

x   auxiliary or concomitant variable available on each unit

z y/π (see below for definition of π)

4. <u>Totals, Means, and their Estimates</u>

T_x, T_y population totals of the variables x and y

$\bar{X}'$, $\bar{Y}'$ population standardized means of the variables x and y

$\hat{T}_y$ Horvitz-Thompson estimator of the population total of the variable y

$$S_{y,A}=\sum_{\mathcal{A}}(y_i-y_j)^2/2(n-1)N$$

$$S_{y,A'}=\sum_{\mathcal{A}'}(y_i-y_j)^2/2(N-n)N$$

$$W_z=\sum_{i\neq j}^{N}\sum^{N}\pi_{ij}(z_i-z_j)^2/2n(n-1)N(N-1)$$

$$W_{z,A}=\sum_{\mathcal{A}}\gamma_{ij}(z_i-z_j)^2/2n(n-1)N'(N'-1).$$

5. <u>Inclusion Probabilities</u>

π_i inclusion probability

π_{ij} pairwise inclusion probability for random-order systematic sampling

γ_{ij} pairwise inclusion probability for fixed-configuration, systematic sampling

π^{o}_{ij} approximate formula for π_{ij} (Overton, 1985)

6. <u>Variances and Variance Estimators</u>

V_x, V_y population variances of the variables x and y

$V(\hat{T}_y)$ variance of the Horvitz-Thompson estimator

$V^*(\hat{T}_y)$ variance of the Horvitz-Thompson estimator for random-order systematic sampling

v_{HT} Horvitz-Thompson variance estimator

v_{YG} Yates-Grundy variance estimator

v^*_{HT} Horvitz-Thompson variance estimator using random-order π_{ij}'s for fixed-configuration sampling

v^*_{YG} Yates-Grundy variance estimator using random-order π_{ij}'s for fixed-configuration sampling

v^o_{HT} Horvitz-Thompson variance estimator calculated using π^o_{ij}

v^o_{YG} Yates-Grundy variance estimator calculated using π^o_{ij}

v^{s1}_{YG} first-order, Yates-Grundy successive difference estimator

v^{s2}_{YG} second-order, Yates-Grundy successive difference estimator

1.2 Systematic Sampling Inference and Designs

Two methods for selecting a variable probability systematic (hereafter, *vps)* sample exist. In *fixed-configuration* systematic sampling, the universe is sorted by the size of an auxiliary variable x, or by the order in which the units are naturally found, say along a transect. In *random-order* systematic sampling, the universe elements are randomly permuted prior to selecting the sample. Detailed descriptions of these sampling methods are given in Hartley and Rao (1962), and Procedures 1 and 2 of Brewer and Hanif (1983).

For variance estimation, two models of inference are identified. The models will be called random-order and fixed-configuration. For the random-order model, variance is estimated using the pairwise inclusion probabilities of random-order sampling. Variance is estimated using the pairwise inclusion probabilities for fixed-configuration selection under the fixed-configuration inference model.

Either model of inference may be invoked for a fixed-configuration or a random-order systematic sample. Both models use the same first-order inclusion probabilities, π_u's, and the Horvitz-Thompson Theorem (Horvitz and Thompson, 1952) still furnishes the proper variance of the Horvitz-Thompson estimator, $\hat{T}_y$. In fixed-configuration inference some pairwise inclusion probabilities are zero, so the usual Horvitz-Thompson and Yates-Grundy variance estimators are biased. Variance estimation for the random-order model of inference has been studied extensively by Stehman and Overton (1987b, 1989).

If a random-order systematic design is employed, variance estimation under the fixed-configuration inference model is conditional inference, conditioning on the particular configuration of units realized from the random ordering of the universe. The random-order inference model applied to a random-order design would then represent estimation of the unconditional variance, that is, the variance over all possible orderings of the universe.

If a fixed-configuration sample is selected, the fixed-configuration inference model is generally of interest. In some circumstances, however, one may wish to view an observed ordering of the universe as a realization of an essentially random generating mechanism. Then the random-order model of inference would provide an unconditional variance estimator appropriate for the hypothetical population from which the particular observed configuration was generated.

The random-order model of inference is frequently used in practice to obtain an estimate of variance for a fixed-configuration design. The sample is treated as if it had been selected from a randomly ordered universe, and a variance estimator appropriate for random-order sampling is then used. In equal probability sampling, it has long been recognized that if a systematic sample results in a smaller variance than would be obtained from a simple random sample, then the variance estimated by the simple random sample variance formula will overestimate the actual fixed-configuration variance. Similarly, if the variance is greater for systematic sampling relative to simple random sampling, the fixed-configuration variance will be underestimated by the random sample variance formula. We might expect this "see-saw" estimator effect to carry over to variable probability systematic sampling.

1.3 Empirical Studies

Several studies of equal probability systematic sampling from natural populations have examined use of the random-order inference model to estimate the fixed configuration variance. Osborne (1942) used systematically placed lines to estimate vegetation composition of an area. Systematic sampling was 2 to 4 times more efficient than equivalent effort stratified random sampling. Use of random sampling formulas to estimate variance of the systematic samples resulted in overestimation of the fixed-configuration variance. In a study estimating crown area and tree frequency in 5 different forest populations, Payendeh (1970) found systematic sampling more efficient than random sampling. Thus a random-order model variance estimator would overestimate the variance of the systematic sample.

In contrast, Milne (1959) studied the random-order model of variance estimation for equal probability systematic samples from 50 natural populations. He concluded that generally one would not go far wrong treating a centric, systematic area-sample as if it were random. Bourdeau (1953) found no advantage in efficiency of systematic sampling over random sampling in estimating basal area of trees. In an empirical study of natural and artificial populations, Wolter (1985) found that the bias of the simple random sampling variance estimator, used to estimate the fixed-configuration variance, was reasonably small in most populations except those showing a strong trend or stratification effect. Although treating the observed configuration as random may result in only a small bias, this is a point that must be established for any circumstance of application.

1.4 Variance Estimation

Because the random-order model of inference tends to result in overestimation of the variance of a fixed-configuration, systematic sample, an estimator is needed that will reflect the gain in precision achieved by fixed-configuration over random-order sampling. More sophisticated variance estimation models can reduce the problem of overestimation. For one-dimensional systematic sampling, a variance estimator based on the mean square successive difference (von Neuman *et al,* 1941) has been found to adequately express the error of equal probability, systematic samples (Overton, 1964). In the usual formula for the variance of $\hat{T}_y$, the sample variance, s^2, is replaced by the mean square successive difference, $\delta^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(y_i - y_{i+1}\right)^2$. Based on empirical investigation of equal probability, systematic sampling from a variety of populations, Wolter (1985) concluded the estimator constructed with δ^2 was a good, generally applicable estimator of the fixed-configuration variance. Wolter reviews many other variance estimators suggested for equal probability, systematic sampling.

Few empirical studies of variance estimators for fixed-configuration, *vps* sampling exist. Schreuder *et al* (1971) compared the properties of several sampling designs and estimators for three forest populations. Among the variable probability sampling methods studied, *vps* performed favorably and was especially good when the universe was in a favorable order. Variance estimators based on the random-order inference model overestimated the fixed-configuration variance except for the smallest sample sizes studied. Wolter (1985) reported similar results for *vps* sampling from a population of 5634 mobile home dealers. Kuk (1989) empirically compared several of the variance estimators listed by Wolter to variance estimators obtained via a bootstrap procedure. Kuk employed a population model relating the y's to the x's to construct the bootstrap population.

Several approaches for obtaining an unbiased estimator of variance under the fixed-configuration inference model apply equally well to variable probability and equal probability systematic sampling (cf. Overton, 1964; Murthy, 1967; Zinger, 1980). These approaches include: 1) supplementing the sample with random observations, 2) interpenetrating subsamples, and 3) model-based approaches.

1.5 Summary of Contents

In section 2, the basic issues relevant to fixed-configuration variance are described using the Horvitz-Thompson formulation of variance. A detailed comparison of the variance for fixed-configuration and random-order sampling is presented in section 3 by investigating the precision of $\hat{T}_y$ for both equal and variable probability sampling. In section 4, the bias of v_{HT} and v_{YG} as estimators of the fixed-configuration variance is derived when the true, fixed-configuration pairwise inclusion probabilities are used, and also when the random-order pairwise inclusion probabilities are used. Alternatives to v_{HT} and v_{YG} for estimating the fixed-configuration variance are described in Section 5. Again, results from equal probability sampling guide the extension to the variable probability case. The alternative estimators are then compared in Section 6 to variance estimators reviewed by Wolter (1985). Empirical comparisons of the fixed-configuration variance to the random-order variance are presented in Section 7. An empirical comparison of variance estimators is also described.

Systematic sampling is examined in a general probability sampling framework, so the results provide a different perspective on the variance estimation problem than is generally obtained from the standard, population model treatment of the problem. The probability sampling approach permits convenient extension of results from equal probability sampling to variable probability sampling. In particular, alternative variance estimators are developed combining fundamental principles of variable probability sampling with well-known equal probability, systematic sampling results. Although the properties of estimators for equal probability, systematic samples are determined by the ordering of the y's, in variable probability sampling these properties are not simply a function of the ordering of the variable $z = y/\pi$. The variable probability structure complicates the relationship between ordering of the z's and properties of the estimators in such a way that prediction of results based only on the ordering of the z's is difficult.

2. VARIANCE ESTIMATION

The general variance estimation issues can be elaborated using the Horvitz-Thompson variance formulation. Variance under the fixed-configuration model is given by

$$V(\hat{T}_y) = \sum_{i=1}^{N} \frac{y_i^2}{\pi_i}(1-\pi_i) + \sum_{\mathcal{A}} (\gamma_{ij} - \pi_i\pi_j) z_i z_j - \sum_{\mathcal{A}'} y_i y_j \tag{1}$$

$$= A + B - C. \tag{2}$$

Because the pairs in the set $\mathcal{A}'$ cannot appear in a sample, the component of variance labelled C in (2) is not estimable from the sample by standard probability sampling methods. The usual form of the Horvitz-Thompson variance estimator,

$$v_{HT} = \sum_{i=1}^{n}{}_{s} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + \sum_{\mathcal{A}}{}_{s} \frac{(\gamma_{ij} - \pi_i\pi_j)}{\gamma_{ij}} z_i z_j ,$$

is unbiased for A+B of the fixed-configuration variance (2). Thus the bias of v_{HT} is $C=\sum_{\mathcal{A}'} y_i y_j$. In addition to being biased, v_{HT} may be unstable because some of the γ_{ij}'s may be extremely small.

Several approaches to estimating the fixed-configuration variance could be pursued. The bias term, C, could be estimated using information in the sample and a prediction model based on some knowledge of the population order. The stability of the estimate of component B could be improved by making the very small γ_{ij}'s slightly larger, that is, by "scoring" the smallest γ_{ij}'s to some pre-set minimum value. Scoring may result in a potentially substantial reduction in mean square error at the expense of small bias. Both of these approaches merit further study, but they will not be considered further.

A common approach to estimating the fixed-configuration variance is to use the random-order model of inference. The variance estimator under this model is

$$v^*_{HT}=\sum_{i=1}^{n} \frac{y_i^2}{\pi_i^2}(1-\pi_i)+\sum_{\mathcal{A}} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} z_i z_j .$$

v^*_{HT} has two components of bias. The component C of equation (2) is still not estimable, and component B is estimated with bias because the π_{ij}'s are not the appropriate fixed-configuration pairwise inclusion probabilities. Both bias components must be considered simultaneously.

Summarizing the main issues of estimating the fixed-configuration variance:

1) the component of variance contained in the summation over $\mathcal{A}'$ is not estimable;
2) some of the fixed-configuration pairwise inclusion probabilities may be very small, thus making the variance estimator highly unstable;
3) variance estimators using the random-order π_{ij}'s will be more stable than the estimators using the true, fixed-configuration γ_{ij}'s.

3. COMPARISON OF RANDOM-ORDER AND FIXED-CONFIGURATION VARIANCES

3.1 Equal Probability Sampling

The relative precision of fixed-configuration sampling versus random-order sampling depends on the fixed-configuration and random-order pairwise inclusion probabilities. For equal probability sampling, assume, for simplicity, that the sampling interval, $k=N/n$, is an integer. Then for fixed-configuration sampling, $\gamma_{ij}=n/N$ if the pair (i,j) can appear in the same sample, $\gamma_{ij}=0$ otherwise. For random-order sampling, $\pi_{ij}=n(n-1)/N(N-1)$ for all pairs (i,j). There are N(n-1) pairs in the set $\mathcal{A}$ for equal probability, fixed-configuration sampling, and N(N-n) pairs for which $\gamma_{ij}=0$ in the set $\mathcal{A}'$. The set of all pairs in the universe is partitioned by $\mathcal{A}$ and $\mathcal{A}'$.

The fixed-configuration variance is given by:

<u>Result 3.1</u> $V(\hat{T}_y) = \frac{1}{2}(1-1/f)\sum_{\mathcal{A}}(y_i-y_j)^2 + \frac{1}{2}\sum_{\mathcal{A}'}(y_i-y_j)^2.$

<u>Proof:</u> This result follows from the more general formula given subsequently in (6), and setting π_i=n/N, and γ_{ij}=n/N, the values appropriate for equal probability sampling. □

To facilitate comparison of fixed-configuration and random-order sampling, note that V_y can be decomposed into two components as follows:

$$\begin{aligned} V_y &= \frac{1}{2N(N-1)} \sum_{i \neq j}^{N}\sum^{N}(y_i-y_j)^2 \\ &= \frac{1}{2N(N-1)} \left[\sum_{\mathcal{A}}^{N}(y_i-y_j)^2 + \sum_{\mathcal{A}'}^{N}(y_i-y_j)^2\right] \\ &= \frac{1}{2N(N-1)} \left[2(n-1)NS_{y,A} + 2(N-n)NS_{y,A'}\right] \\ &= \left(\frac{n-1}{N-1}\right)S_{y,A} + \left(\frac{N-n}{N-1}\right)S_{y,A'}, \end{aligned} \qquad (3)$$

where,

$$S_{y,A} = \sum_{\mathcal{A}}(y_i-y_j)^2/2(n-1)N,$$

and

$$S_{y,A'} = \sum_{\mathcal{A}'}(y_i-y_j)^2/2(N-n)N.$$

Then Result 3.1 can be re-written as

$$V(\hat{T}_y) = N(N-1)V_y - N^2\left(\frac{n-1}{n}\right)S_{y,A}. \qquad (4)$$

Comparing the fixed-configuration variance to the random-order variance, we obtain:

<u>Result 3.2</u> $V^*(\hat{T}_y) - V(\hat{T}_y) = N^2\left(\frac{n-1}{n}\right)\left[S_{y,A} - V_y\right]$ (5)

$$= N^2\left(\frac{N-n}{N-1}\right)\left[S_{y,A} - S_{y,A'}\right]$$

<u>Proof</u>: Using (4) and the result $V^*(\hat{T}_y) = N^2\left(\frac{N-n}{N}\right)\frac{V_y}{n}$, equation (5) follows after some algebra. □

Result 3.2 leads to the following theorem:

<u>Theorem</u> <u>3.3</u> If $S_{y,A} > V_y$, then $V^*(\hat{T}_y) > V(\hat{T}_y)$.

<u>Proof</u>: From (5), if $N^2\left(\frac{n-1}{n}\right)\left[S_{y,A} - V_y\right] > 0$, then $V^*(\hat{T}_y) > V(\hat{T}_y)$. Since $N^2\left(\frac{n-1}{n}\right) > 0$, the theorem follows. □

Theorem 3.3 implies that fixed-configuration sampling is more efficient than random-order sampling when $S_{y,A} > V_y$. Because $S_{y,A}$ is equivalent to Cochran's (p. 207, 1977) within systematic sample variance, $S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$, Theorem 3.3 is a restatement of Cochran's Theorem 8.1. Theorem 3.3 shows that systematic sampling is beneficial when the large differences in the population, $(y_i - y_j)$, are allocated to subset $\mathcal{A}$; that is, the large differences should be associated with the non-zero pairwise inclusion probabilities to maximize $S_{y,A}$. This strategy agrees with the common recommendation of maximizing S^2_{wsy} in order to obtain high precision with an equal probability, systematic sample.

3.2 Variable Probability Sampling

The fixed-configuration variance is given by $V(\hat{T}_y) = \frac{1}{2} \sum_{j \neq i}^{N} \sum^{N} (\pi_i \pi_j - \gamma_{ij})(z_i - z_j)^2$. Since $\gamma_{ij} = 0$ for all $(i,j) \varepsilon \mathcal{A}'$, we obtain

$$V(\hat{T}_y) = \frac{1}{2} \sum_{\mathcal{A}} (\pi_i \pi_j - \gamma_{ij})(z_i - z_j)^2 + \frac{1}{2} \sum_{\mathcal{A}'} \pi_i \pi_j (z_i - z_j)^2. \tag{6}$$

The corresponding variance for random-order sampling is

$$V^*(\hat{T}_y) = \frac{1}{2}\sum_{\mathcal{A}}(\pi_i\pi_j-\pi_{ij})(z_i-z_j)^2 + \frac{1}{2}\sum_{\mathcal{A}'}(\pi_i\pi_j-\pi_{ij})(z_i-z_j)^2.$$

Then the difference between the fixed-configuration and random-order variances is

$$V^*(\hat{T}_y) - V(\hat{T}_y) = \frac{1}{2}\sum_{\mathcal{A}}(\gamma_{ij}-\pi_{ij})(z_i-z_j)^2 - \frac{1}{2}\sum_{\mathcal{A}'}\pi_{ij}(z_i-z_j)^2. \tag{7}$$

Fixed-configuration sampling will have smaller variance than random-order sampling if the large differences (z_i-z_j) are allocated to $\mathcal{A}$. But in contrast to equal probability, systematic sampling, variable probability sampling requires these large differences to be associated with positive $(\gamma_{ij}-\pi_{ij})$. That is, the squared difference (z_i-z_j) must be generally large, and the pairwise inclusion probability of this (i,j) pair under fixed-configuration sampling must be larger than the corresponding pairwise inclusion probability under random-order sampling.

Properties of the estimators in equal probability, systematic sampling depend on the ordering of the y's. In variable probability sampling, properties depend on the ordering of the z's and the relationship of this ordering to the x's. For example, if the universe is sorted on the y's, equal probability, systematic sampling is much more efficient than simple random sampling. The analogous situation in variable probability sampling would be to sort the universe on z. In general, fixed-configuration systematic sampling from this ordering of the z's should be more efficient than random-order sampling. But it is possible that for some configurations of the x's, the γ_{ij}'s will be such that random-order sampling is more efficient than fixed-configuration sampling. It is difficult to predict the relationship between γ_{ij} and π_{ij} under the two *vps* designs. For example, if π_k and π_l are both large, the random-order π_{kl} will be relatively large compared to all other π_{ij}'s. Under fixed-configuration sampling, for π_k and π_l large, γ_{kl} could be very small, even zero.

The relationship of the pairwise inclusion probabilities and the z's can be illustrated further by the development of a variable probability analog to Theorem 3.3. Upon rewriting equation (7) as

$$V^*(\hat{T}_y)-V(\hat{T}_y)=\frac{1}{2}\sum_{\mathcal{A}}\gamma_{ij}(z_i-z_j)^2-\frac{1}{2}\sum_{i\neq j}^{N}\sum^{N}\pi_{ij}(z_i-z_j)^2, \tag{8}$$

the quantities in the two double summations in (8) are recognized as analogs to V_y and $S_{y,A}$. Define $V_z=\sum_{i\neq j}^{N}\sum^{N}(z_i-z_j)^2/2N(N-1)$, and $S_{z,A}=\sum_{\mathcal{A}}(z_i-z_j)^2/2N'(N'-1)$, where N' is the number of pairs in the universe for which $\gamma_{ij}=0$. If the ordering of the z's were the lone determining factor in the comparison of fixed-configuration to random-order sampling, the variable probability analog of Theorem 3.3 would be obtained by substituting z for y in Theorem 3.3, and decomposing V_z into components $S_{z,A}$ and $S_{z,A'}$. But the true variable probability relationship depends instead on a decomposition weighted by the γ_{ij}'s and π_{ij}'s. Note that

$\sum_{\mathcal{A}}\gamma_{ij}=\sum\sum_{i\neq j}^{N\,N}\pi_{ij}=n(n-1)$. Then define

$$W_z=\frac{1}{2N(N-1)}\sum\sum_{i\neq j}^{N\,N}\pi_{ij}(z_i-z_j)^2/\sum\sum_{i\neq j}^{N\,N}\pi_{ij}$$

$$=\sum\sum_{i\neq j}^{N\,N}\pi_{ij}(z_i-z_j)^2/2n(n-1)N(N-1),\text{ and}$$

$$W_{z,A}=\frac{1}{2N'(N'-1)}\sum_{\mathcal{A}}\gamma_{ij}(z_i-z_j)^2/\sum_{\mathcal{A}}\gamma_{ij}$$

$$=\sum_{\mathcal{A}}\gamma_{ij}(z_i-z_j)^2/2n(n-1)N'(N'-1).$$

Theorem 3.4 If $W_{z,A}>\frac{N(N-1)}{N'(N'-1)}W_z$, then $V^*(\hat{T}_y)>V(\hat{T}_y)$.

Proof: Using the definitions above, write equation (8) as $V^*(\hat{T}_y)-V(\hat{T}_y)=n(n-1)\left[N'(N'-1)W_{z,A}-N(N-1)W_z\right]$, and the theorem follows directly.

Theorem 3.4 is the variable probability analog of Theorem 3.3

4. VARIANCE ESTIMATION USING CONVENTIONAL FORMULAS

The biases of the Horvitz-Thompson and Yates-Grundy variance estimators are investigated for fixed-configuration sampling. Bias results for the general case of *vps* sampling are established first, and results for equal probability sampling are derived later as special cases. The biases for the variance estimators using the fixed-configuration γ_{ij}'s are:

Result 4.1

i) $\text{Bias}(v_{HT})=\sum_{\mathcal{A}'}y_iy_j$.

ii) $\text{Bias}(v_{YG})=-\frac{1}{2}\sum_{\mathcal{A}'}\pi_i\pi_j(z_i-z_j)^2$.

Proof:

i) Follows from equations (1) and (2) of section 2.

ii) $E(v_{YG}) = E\left[\frac{1}{2}\sum_{i\neq j}^{n}\sum^{n}\left(\frac{\pi_i\pi_j}{\gamma_{ij}}-1\right)(z_i-z_j)^2\right]$

$$=\frac{1}{2}\sum_{\mathcal{A}}\left(\frac{\pi_i\pi_j}{\gamma_{ij}}-1\right)\gamma_{ij}(z_i-z_j)^2$$

$$=\frac{1}{2}\sum_{\mathcal{A}}\left(\pi_i\pi_j-\gamma_{ij}\right)\gamma_{ij}(z_i-z_j)^2. \qquad (10)$$

Subtracting (6) from (10) yields (9). □

Result 4.1 ii) was apparently also recognized by Brewer and Hanif (p. 22, 1983).

For the random-order model of inference, which uses the π_{ij}'s in place of the γ_{ij}'s, the biases of the variance estimators are:

Result 4.2

i) $\text{Bias}(v^*_{HT}) = \sum_{i\neq j}^{N}\sum^{N} y_i y_j - \sum_{\mathcal{A}}\frac{\gamma_{ij}}{\pi_{ij}}y_i y_j.$

ii) $\text{Bias}(v^*_{YG}) = \frac{1}{2}\sum_{\mathcal{A}}\pi_i\pi_j\left(\frac{\gamma_{ij}}{\pi_{ij}}-1\right)(z_i-z_j)^2-\frac{1}{2}\sum_{\mathcal{A}'}\pi_i\pi_j(z_i-z_j)^2. \qquad (11)$

Proof:

i) $E(v^*_{HT}) = \sum_{i=1}^{N}\frac{y_i^2}{\pi_i}(1-\pi_i)+\sum_{\mathcal{A}}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\gamma_{ij}z_i z_j. \qquad (12)$

Subtracting (1) from (12) gives the result.

ii) $E(v^*_{YG}) = E\left[\frac{1}{2}\sum_{i\neq j}^{n}\sum^{n}\left(\frac{\pi_i\pi_j}{\pi_{ij}}-1\right)(z_i-z_j)^2\right]$

$$=\frac{1}{2}\sum_{i\neq j}^{N}\sum^{N}\left(\frac{\pi_i\pi_j}{\pi_{ij}}-1\right)\gamma_{ij}(z_i-z_j)^2$$

$$=\frac{1}{2}\sum_{\mathcal{A}}\left(\frac{\pi_i\pi_j}{\pi_{ij}}-1\right)\gamma_{ij}(z_i-z_j)^2 \quad \text{since } \gamma_{ij}=0 \text{ for } \mathcal{A}'. \qquad (13)$$

Subtracting (6) from (13) leads to (11). □

Biases of the variance estimators for equal probability sampling are readily derived from Results 4.1 and 4.2. For the estimators calculated using the fixed-configuration γ_{ij}'s:

Result 4.3

i) $\text{Bias}(v_{HT}) = \sum_{\mathcal{A}'} y_i y_j$.

ii) $\text{Bias}(v_{YG}) = -\frac{1}{2}\sum_{\mathcal{A}'} (y_i - y_j)^2$ (14)

$$= (n-1)NS_{y,A} - N(N-1)V_y. \quad (15)$$

Proof:

i) Follows directly from Result 4.1 i).

ii) Equation (14) follows by substituting the appropriate equal probability values for π and z into (9). After some algebra requiring (3), we obtain (15). □

The bias of v_{HT} is $\sum_{\mathcal{A}'} y_i y_j$, whether sampling is with equal or unequal probabilities, so for a non-negative response variable y, the bias of v_{HT} is always positive. For equal probability sampling, v_{YG} is negatively biased. Note then, that for equal probability sampling, the Horvitz-Thompson and Yates-Grundy variance estimators are equal for random-order sampling, but differ from each other under fixed-configuration sampling.

For equal probability sampling and the random-order inference model, we have:

<u>Result 4.4</u>

i) $\text{Bias}(v^*_{HT}) = \sum_{i \neq j}^{N}\sum^{N} y_i y_j - \frac{N-1}{n-1}\sum_{\mathcal{A}} y_i y_j$. (16)

ii) $\text{Bias}(v^*_{YG}) = N(N-1)\left[S_{y,A} - V_y\right]$. (17)

<u>Proof</u>:

i) Equation (16) is obtained by substituting $\gamma_{ij} = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$ into Result 4.2 i).

ii) Equation (17) follows from Result 4.2 ii) using $\pi_{ij} = n/N$ and the definitions of V_y and $S_{y,A}$. □

From (17), the "see-saw" variance estimator effect of the random-order model is evident. As $S_{y,A}$ increases, the efficiency of systematic sampling relative to simple random sampling increases, and the positive bias of v^*_{YG} increases. Since $v^*_{YG} = v^*_{HT}$ under this equal probability, random-order model, v^*_{HT} would also display the see-saw effect.

The see-saw variance estimator effect appears to be diminished in variable probability sampling. Repeating equation (11),

$$\text{Bias}(v^*_{YG}) = \frac{1}{2}\sum_{\mathcal{A}} \pi_i \pi_j \left(\frac{\gamma_{ij}}{\pi_{ij}} - 1\right)(z_i - z_j)^2 - \frac{1}{2}\sum_{\mathcal{A}'} \pi_i \pi_j (z_i - z_j)^2.$$

The components of the see-saw effect are present if the term $\left(\frac{\gamma_{ij}}{\pi_{ij}} - 1\right)$ is ignored. If the large differences $(z_i - z_j)$ are in $\mathcal{A}$, so that fixed-configuration, systematic sampling is more efficient than random-order, *vps* sampling (see equation (8)), then v^*_{YG} would be positively biased. But the terms $\left(\frac{\gamma_{ij}}{\pi_{ij}} - 1\right)$ will decrease this positive bias. Usually $\left(\frac{\gamma_{ij}}{\pi_{ij}} - 1\right)$ will be positive because $\gamma_{ij} > \pi_{ij}$ for most non-zero γ_{ij}. But if a substantial proportion of the $\left(\frac{\gamma_{ij}}{\pi_{ij}}-1\right)$ are negative, the positive bias of v^*_{YG} could be small. Thus some tendency toward a see-saw effect should be observed, but the effect will not be as strong as it is in equal probability, systematic sampling. Similar interpretations would be obtained for v^*_{HT}. The empirical results of Section 7 verify the absence of a strong see-saw effect for v^*_{YG}.

5. ALTERNATIVE VARIANCE ESTIMATORS

When fixed-configuration sampling results in smaller variance than random-order sampling, the estimators v^*_{HT} and v^*_{YG} will often overestimate variance. In this section, variance estimators are developed analogous to the equal probability sampling variance estimator based on the mean square successive difference,

$$\hat{V}(\hat{T}_y) = \frac{N^2(1-f)}{2n(n-1)} \sum_{i=1}^{n-1} (y_i - y_{i+1})^2 . \tag{18}$$

The estimator in (18) is similar in form to the Yates-Grundy variance estimator, so the development of the fixed-configuration variance estimators proceeds naturally from the Yates-Grundy form.

A family of variance estimators can be constructed by partitioning the Yates-Grundy variance estimator. Rewrite v^*_{YG} as a linear function of a set of difference estimators, each with different order lag (Stehman and Overton, 1987):

$$v^*_{YG} = \sum_{j=1}^{n-1} a_j v^{*j}_{YG},$$

where
$$v_{YG}^{sj} = b_j \sum_{i=1}^{n-j} c_{ij} d_{ij}^2, \tag{19}$$

$$c_{ij} = (\pi_i \pi_{i+j} - \pi_{i,i+j}) / \pi_{i,i+j}, \text{ and}$$

$$d_{ij} = \left[\frac{y_i}{\pi_i} - \frac{y_{i+j}}{\pi_{i+j}}\right].$$

In this formulation, $a_j = b_j^{-1}$. Then v_{YG}^{sj} denotes a family of estimators, where j is specified to provide an estimator of lag j.

Consider the first-order successive differences, $d_{i1} = \left[\frac{y_i}{\pi_i} - \frac{y_{i+1}}{\pi_{i+1}}\right]$. The resultant formula for the first order variance estimator is:

$$v_{YG}^{s1} = \frac{n}{2} \sum_{i=1}^{n-1} \left(\frac{\pi_i \pi_{i+1}}{\pi_{i,i+1}} - 1\right)\left(\frac{y_i}{\pi_i} - \frac{y_{i+1}}{\pi_{i+1}}\right)^2. \tag{20}$$

The value $b_1 = n/2$ in (20) was chosen to scale the variance estimator to a sum over n(n-1)/2 differences. The usual Yates-Grundy estimator is a sum of n(n-1)/2 terms, but only n-1 differences are summed in v_{YG}^{s1}. The multiplier n/2 in v_{YG}^{s1} scales the estimator to be equivalent to a sum over n(n-1)/2 terms. If sampling is with uniform inclusion probabilities, v_{YG}^{s1} reduces to the estimator in (18). This further justifies choice of n/2 as the scaling factor.

What to use for $\pi_{i,i+1}$ is open to question. The random-order π_{ij}'s are used in equation (18) for equal probability sampling. The good performance of this estimator in empirical studies suggests that in variable probability sampling, the random-order π_{ij}'s should also be used, instead of the fixed-configuration γ_{ij}'s. Subsequently, when calculating v_{YG}^{s1}, $\pi_{i,i+1}$

will be replaced by $\pi^{o}_{i,i+1}$, an approximation to the random-order, *vps* sampling pairwise inclusion probabilities, where $\pi^{o}_{i,i+1}=\frac{2(n-1)\pi_i\pi_{i+1}}{2n-\pi_i-\pi_{i+1}}$ (Overton, 1985). The random-order approximation of Hartley and Rao (1962) could also be used instead of $\pi^{o}_{i,i+1}$.

The second-order, Yates-Grundy successive difference estimator was also investigated in the simulation study described in Section 7. The specific formula for this estimator is:

$$v^{s2}_{YG}=\frac{n(n-1)}{2(n-2)}\sum_{i=1}^{n-2}\left(\frac{\pi_i\pi_{i+2}}{\pi_{i,i+2}}-1\right)\left(\frac{y_i}{\pi_i}-\frac{y_{i+2}}{\pi_{i+2}}\right)^2 .$$

Again the scaling factor, here $b_2=n(n-1)/2(n-2)$, was chosen to make the estimator equivalent to a sum over $n(n-1)/2$ squared differences.

More complex scaling factors could easily be envisioned. A scaling that would take into account the magnitude of the $c_{i,i+j}$'s in a successive difference estimator relative to the c_{kl}'s in the entire sample would be $b_j=\sum\sum_{k\neq l} c_{kl}/\sum_{i=1}^{n-j} c_{i,i+j}$. For example, if the $c_{i,i+1}$'s used in v^{s1}_{YG} were the small values among the c_{kl}'s in the entire sample, the scaling factor b_1 would be larger to compensate for the small multipliers on the squared differences, d^2_{i1}. For equal probability sampling, this more complicated scaling is equivalent to the scaling described earlier. Simulation results showed only minor differences in the variance estimators using the different scalings, so the simpler scaling procedure was retained.

6. COMPARISON OF ALTERNATIVE VARIANCE ESTIMATORS

Wolter (Chapter 7, 1985) provides an extensive investigation of variance estimation for fixed-configuration, *vps* sampling. Several variance estimators are proposed, most of which are derived as variable probability analogs to common variance estimators from equal probability sampling. The equal probability sampling variance estimators are derived on the basis of an hypothesized ordering of the y's in fixed-configuration.

Two estimators of the fixed-configuration variance are motivated by treating a systematic sample as approximately a random stratified sample with one unit per stratum (cf. Hendricks, 1956). If the differences in the means of adjacent strata are small, it is reasonable to "collapse" to n/2 strata, and regard the systematic sample as if $n_h=2$ units had been selected at random from each of the n/2 strata. Wolter's variance estimators

$$v_{11}=\frac{1}{n^2}\sum_{i=1}^{n/2}{}_s\left(\frac{y_{2i}}{p_{2i}}-\frac{y_{2i-2}}{p_{2i-2}}\right)^2$$

and

$$v_{12}=\frac{1}{2n(n-1)}\sum_{i=2}^{n}{}_s\left(\frac{y_i}{p_i}-\frac{y_{i-1}}{p_{i-1}}\right)^2,$$

where $p_i=x_i/T_x$, are derived from the stratified sampling approximation. Non-overlapping differences are used in v_{11}, while degrees of freedom are increased by allowing overlapping differences in v_{12}. In order to account for the without replacement feature of the sampling design, Wolter (1985) constructs two additional estimators by pre-multiplying v_{11} and v_{12} by an approximate finite population correction ($\hat{fpc}$), to obtain $v_{14}=(\hat{fpc})v_{11}$, and $v_{15}=(\hat{fpc})v_{12}$. The approximate fpc used is $1-\sum_{i=1}^{n}\pi_i/n=1-\bar{\pi}_s$. The choice of an approximate fpc is not unique. For example, Yates and Grundy (1953) suggested the approximation $(1-\sum_{i=1}^{N}\pi_i^2/n)=1-E(\bar{\pi}_s)$. If π_i=n/N is substituted into these approximate fpc's, the result is 1-f.

The estimator $v_{YG}^{\#1}$ was described earlier as a variable probability analog to the mean square successive difference estimator. $v_{YG}^{\#1}$ can also be motivated by the stratified sampling approximation argument. Let π_i and π_j represent the inclusion probabilities for any two units in the fixed-configuration, *vps* sample. If a random-order, *vps* sample of size 2 is selected within each stratum such that the inclusion probabilities are still π_i and π_j, then π_{ij}^o for any 2 units within a stratum is numerically equivalent to π_{ij}^o for the same 2 units when a random-order, *vps* sample is selected from the entire population. Thus, the random-order approximation π_{ij}^o is appropriate for within stratum pairwise inclusion probabilities in this stratified sampling model. Since $\pi_{ij}=0$ for elements i and j in different strata, $v_{YG}^{\#1}$ is an appropriate stratified sampling variance estimator, with π_{ij}^o substituted for π_{ij} for the non-zero, within stratum pairwise inclusion probabilities. An advantage of this derivation of

v_{YG}^{s1} over v_{14} and v_{15} is that the without replacement feature of the design is incorporated into the variance estimator, and not appended at the end as in v_{14} and v_{15}.

Clearly, v_{YG}^{s1}, v_{15}, and v_{12} are closely related. Since $v_{15} = (f\hat{p}c)\, v_{12}$, $v_{15} < v_{12}$. Further,

$$v_{YG}^{s1} = v_{12} - \sum_{i=1}^{n-1} \left(\frac{\pi_i + \pi_{i+1}}{2}\right)\left(\frac{y_i}{\pi_i} - \frac{y_{i+1}}{\pi_{i+1}}\right)^2,$$

so $v_{YG}^{s1} \leq v_{12}$. No general inequality holds for v_{YG}^{s1} and v_{15}, but the nature of the difference is apparent by re-writing v_{15} as

$$v_{15} = \frac{n}{2(n-1)} \sum_{i=1}^{n-1} \left(1 - \frac{\bar{\pi}_s + \bar{\pi}_s}{2}\right)\left(\frac{y_i}{\pi_i} - \frac{y_{i+1}}{\pi_{i+1}}\right)^2, \tag{21}$$

and noting that

$$v_{YG}^{s1} = \frac{n}{2(n-1)} \sum_{i=1}^{n-1} \left(1 - \frac{\pi_i + \pi_{i+1}}{2}\right)\left(\frac{y_i}{\pi_i} - \frac{y_{i+1}}{\pi_{i+1}}\right)^2. \tag{22}$$

v_{15} may then be viewed as a modification of v_{YG}^{s1} in which $\pi_{i,i+1}$ is replaced by an approximation similar in form to $\pi_{i,i+1}^{o}$, but with π_i and π_j replaced by $\bar{\pi}_s$. That is, in v_{15}, $\pi_{i,i+1}$ is approximated by $\hat{\pi}_{i,i+1} = \frac{2(n-1)\bar{\pi}_s\bar{\pi}_s}{2n - \bar{\pi}_s - \bar{\pi}_s}$. Only minor differences between the properties of v_{15} and v_{YG}^{s1} were observed in the empirical comparison of the estimators.

7. EMPIRICAL RESULTS

A simulation study, similar in design to the population space analysis of Stehman and Overton (1989), was conducted to explore properties of variance and variance estimators in fixed-configuration, *vps* sampling. The two families of populations investigated, GAMNORM and BIGAMMA, were generated from known probability distributions. Within each family, three different subfamilies representing low, medium, and high correlations between the response variable, y, and the design covariate, x, were studied. A subfamily was then a set of populations with the same correlation, within the same major family.

All populations within a subfamily were created from a single base population. A subfamily was created from the base population by adding or subtracting constants to x and/or y. Thus all populations in a subfamily are the same "cloud" of points shifted to various locations in the (x,y)-plane. The variables x and y were standardized, $X'=x/\sqrt{V_x}$ and $Y'=y/\sqrt{V_y}$, and the standardized population centroid, $(\bar{X}', \bar{Y}')$, was used to locate populations within the population space. All populations within a subfamily have $V_y=V_x$, where V_x and V_y are the population variances of x and y, respectively, and populations within a subfamily also have the same V_y and the same correlation between x and y. Populations differing by an additive shift in the x's have different inclusion probabilities.

For the GAMNORM family of populations, x was randomly generated from a standard gamma distribution with parameters $\alpha=2$ and $\lambda=1$, and y was generated, conditional on x, as a normal random variable. For each x_i, y_i was obtained from the equation, $y_i=\beta x_i+\epsilon_i$, where ϵ_i was a random variable distributed Normal$(0, \sigma_\epsilon^2)$, and $\sigma_\epsilon^2=(1-\rho^2)\sigma_x^2$. The same set of 100 x's was used as the base population for all three GAMNORM subfamilies.

The BIGAMMA family was generated from the bivariate gamma distribution with the marginal distributions both standard gamma with parameter $\alpha=2$. For the BIGAMMA family, a different set of x's was generated for each of the subfamily base populations.

The random-order variance estimators v^o_{HT} and v^o_{YG}, and four other variance estimators designed for use in fixed-configuration sampling, denoted by v^{s1}_{YG}, v^{s2}_{YG}, v_{11}, and v_{15} were investigated. The properties of the alternative variance estimators were investigated for random-order *vps* sampling, fixed-configuration *vps* sampling sorting the universe on x (fixed-order(x)), and fixed-configuration *vps* sampling sorting the universe on z (fixed-order(z)). Sample size was n=16, and results were based on 5,000 repetitions of the sampling procedure. The simulation programs were written in the GAUSS programming language (Aptech Systems, Inc., Kent, Washington).

The results for fixed-configuration sampling differed from the random-order, *vps* results reported in Stehman and Overton (1989) in that the fixed-configuration surfaces were more complex. Significant differences in the fixed-configuration behavior surfaces were also evident between the two families examined. Providing general conclusions for fixed-configuration sampling was hampered because of the irregularities of these behavior surfaces.

7.1 Comparison of Variance

The variance of $\hat{T}_y$ for the fixed-configuration systematic designs was compared to $V(\hat{T}_y)$ for random-order, *vps* sampling. Figure 1 (all figures appear following Section 8) shows the surfaces for the ratio of the variance for fixed-order(x) sampling relative to random-order sampling. The main features of these surfaces were:

1) Generally, fixed-order(x) sampling was more efficient than random-order sampling, but exceptions to this rule occurred;
2) For the low correlation subfamilies, the surfaces increased from the upper left corner of the population space toward the lower right corner; fixed-order(x) was more efficient in the upper left corner, while random-order sampling was more efficient in the lower right corner;
3) Fixed-order(x) sampling was more efficient than random-order sampling over the entire population space for the GAMNORM75 subfamily, but random-order sampling was more efficient for a large region of the BIGAMMA77 subfamily;
4) For the high correlation subfamilies, little advantage was gained by fixed-order(x) sampling along the standard diagonal of the GAMNORM subfamily, but fixed-order(x) was more efficient than random-order sampling for most of the BIGAMMA subfamily; a ridge extended along the standard

diagonal for both high correlation subfamilies.

Figure 2 shows the ratio of the variance of fixed-order(z) to random-order *vps* sampling. General conclusions for this figure were:

1) Fixed-order(z) sampling was more efficient than random-order sampling, with few exceptions, and the gain in precision of fixed-order(z) was sometimes substantial;
2) Random-order sampling was more efficient than fixed-order(z) along the standard diagonal for the high correlation subfamilies, the lower right corner of GAMNORM48, and one population in BIGAMMA77.

The exceptions to the general patterns observed in the simulations verified the theoretical results derived earlier. Certain configurations of the x's may result in γ_{ij}'s that make random-order sampling more efficient than fixed-order(z) sampling. Sorting on the ratios, z, usually resulted in better precision, but did not guarantee increased precision, relative to random-order sampling.

The ratio of the variance of fixed-order(x) to the variance of fixed-order(z) is shown in Figure 3:

1) The surfaces for the BIGAMMA family were very different from the GAMNORM family surfaces because the BIGAMMA family had greater differences between the two fixed-configuration selection procedures;
2) Fixed-order(x) and fixed-order(z) were nearly equal for

GAMNORM75 and GAMNORM94, with the exception of one population;

3) Fixed-order(z) was more efficient than fixed-order(x) for most of the BIGAMMA49 and BIGAMMA77 population space;
4) For the BIGAMMA97 subfamily, sorting on x was more efficient than sorting on z for populations above the standard diagonal, but less efficient for populations below the standard diagonal.

7.2 Estimators of the Fixed-configuration Variance

Given the complicated behaviors of the fixed-configuration variances, it is not surprising that the variance estimators have similarly complex behavior surfaces. Simulation results are presented only for v^{o}_{YG}, v^{*1}_{YG}, and v^{*2}_{YG}. v_{15} and v^{*1}_{YG} were remarkably similar in their behavior, as predicted from equations (21) and (22), so results shown for v^{*1}_{YG} apply also to v_{15}. v_{11} significantly overestimated the random-order variance and was eliminated from consideration. The relationship between v^{o}_{HT} and v^{o}_{YG} in fixed-configuration sampling was the same as observed in random-order, *vps* sampling (Stehman and Overton, 1989). The estimators v_{15}, v^{*1}_{YG}, and v^{*2}_{YG} performed similarly to v^{o}_{YG} in random-order sampling, so the three fixed-configuration variance estimators were adequate for random-order inference.

Relative bias results are plotted in Figures 4 and 5 for fixed-order(x) sampling. Relative bias showed no discernible pattern for v_{YG}^{*1} (and, therefore, also v_{15}) and v_{YG}^{*2}. For v_{YG}^{o}, however, relative bias was related to the fixed-configuration variance. The bias was positive if the fixed-configuration variance was less than the random-order variance, and the bias was negative if the fixed-configuration variance was greater than the random-order variance.

This pattern of bias of v_{YG}^{o} was not indicative of a see-saw effect, however. Table 1 illustrates the absence of a see-saw effect using the BIGAMMA family as an example. Columns (3) and (5) in the table show the expected value of v_{YG}^{o} under random-order sampling divided by the expected value of v_{YG}^{o} under fixed-configuration sampling. These ratios are close to 1, demonstrating that v_{YG}^{o} estimates, in expectation, $V(\hat{T}_y)$ for the random-order model of inference, even when fixed-configuration sampling has been done. Columns (4) and (6) of the table show that the variance of random-order sampling differs considerably from the fixed-configuration variance, so the nearly constant ratios in columns (3) and (5) are evidence against a see-saw effect in v_{YG}^{o}.

Coverage properties of the variance estimators were linked to relative bias -- large positive bias resulted in close to 100% coverage, while large negative bias led to very low coverage relative to the 95% nominal level. The results using v_{YG}^{*1} and v_{YG}^{*2} were not encouraging. Both improved on the

Table 1 Investigation of the "See-Saw" Estimator Effect for v^{o}_{YG} in Variable Probability Sampling: BIGAMMA Family

Key to values in columns:

(3) Random-order $E(v^{o}_{YG})$ / Fixed-order(x) $E(v^{o}_{YG})$

(4) Random-order $V(\hat{T}_y)$ / Fixed-order(x) $V(\hat{T}_y)$

(5) Random-order $E(v^{o}_{YG})$ / Fixed-order(z) $E(v^{o}_{YG})$

(6) Random-order $V(\hat{T}_y)$ / Fixed-order(z) $V(\hat{T}_y)$

BIGAMMA49

$\bar{X}'$	$\bar{Y}'$	(3)	(4)	(5)	(6)
2.00	7.00	0.95	3.85	0.96	4.17
2.00	2.00	1.01	0.87	0.97	2.86
7.00	2.00	0.98	1.10	0.99	1.85
12.00	12.00	1.00	0.93	0.96	3.45
7.00	12.00	0.97	2.38	0.96	12.50
7.00	7.00	1.00	1.28	0.95	5.56
12.00	7.00	1.02	0.71	0.99	1.27
3.64	6.30	0.97	3.23	0.96	5.26
5.15	5.15	0.99	1.08	0.96	3.23
6.30	3.64	0.99	0.88	0.95	3.23

BIGAMMA77

$\bar{X}'$	$\bar{Y}'$	(3)	(4)	(5)	(6)
2.00	7.00	0.96	7.69	0.99	0.88
2.00	2.00	1.02	0.88	1.00	1.30
7.00	2.00	1.01	0.75	0.97	2.33
12.00	12.00	1.07	0.50	0.96	2.63
7.00	12.00	1.02	0.82	0.99	1.64
7.00	7.00	1.09	0.40	0.96	3.03
12.00	7.00	1.07	0.48	0.96	3.13
3.64	6.30	0.96	4.00	0.97	2.94
5.15	5.15	1.05	0.46	0.95	3.70
6.30	3.64	1.07	0.46	0.97	3.23

Table 1 (Continued)

BIGAMMA97

$\bar{X}'$	$\bar{Y}'$	(3)	(4)	(5)	(6)
2.00	7.00	0.97	4.55	0.97	3.23
2.00	2.00	1.02	0.87	1.02	1.52
7.00	2.00	0.96	4.76	0.96	10.00
12.00	12.00	0.96	1.79	0.96	1.61
7.00	12.00	0.96	5.00	0.99	1.35
7.00	7.00	0.99	1.49	1.00	0.91
12.00	7.00	0.97	2.44	0.96	10.00
3.64	6.30	0.95	6.67	0.96	3.13
5.15	5.15	1.01	1.28	1.03	1.00
6.30	3.64	0.97	2.63	0.95	12.50

extreme overestimation obtained from the random-order estimators v°_{HT} and v°_{YG}, but the underestimation of v^{s1}_{YG} and v^{s2}_{YG} was extreme in some cases (GAMNORM94, for example), and these estimators cannot be used with confidence at this stage.

8. CONCLUSIONS

The theoretical results presented in this paper help to explain the behavior of the fixed-configuration variance and estimators of this variance. General conclusions are difficult because the fixed-configuration properties depend on the γ_{ij}'s, which are determined by the particular set and configuration of x's in the universe.

The theoretical and empirical results provide some rough guidelines for survey design. Fixed-order(x) was usually a more efficient design than random-order sampling. Along the left edge of the population space, the decrease in variance obtained by fixed-order(x) relative to random-order *vps* sampling was often substantial. Random-order sampling has a large variance relative to simple random sampling in this region. Fixed-order(x) sampling offers an alternative to the strategy of using random-order *vps* sampling and shifting populations out of this region, as suggested in Stehman and Overton (1989). Although sorting on the x's was often an efficient strategy, the unpredictability of the exceptions to this general pattern (e.g. GAMNORM75) prevents recommending this strategy.

The variance estimators investigated in the empirical study were not satisfactory in some populations. Although v^{s1}_{YG} and v^{s2}_{YG} sometimes improved on the random-order variance estimators, v^{s1}_{YG} and v^{s2}_{YG} often had the undesirable property of severely underestimating variance. Wolter's (1985) empirical study showed much better behavior for the fixed-configuration variance estimators, including the estimator v_{15}, which performs similarly to v^{s1}_{YG}. Wolter investigated larger populations and sample sizes, perhaps better conditions for good behavior of the variance estimators. Samples of size 16 may be too small for the fixed-configuration variance estimators to work well. v^{s1}_{YG} performed well for samples of size 30 in the five populations examined by Kuk (1989).

Properties of variance and variance estimators in fixed-configuration *vps* sampling are better understood, but the behaviors of the estimators are complex. A more extensive empirical study would help to determine if there are, in fact, general patterns to these behaviors, or if they are as unpredictable as they appeared in the empirical study.

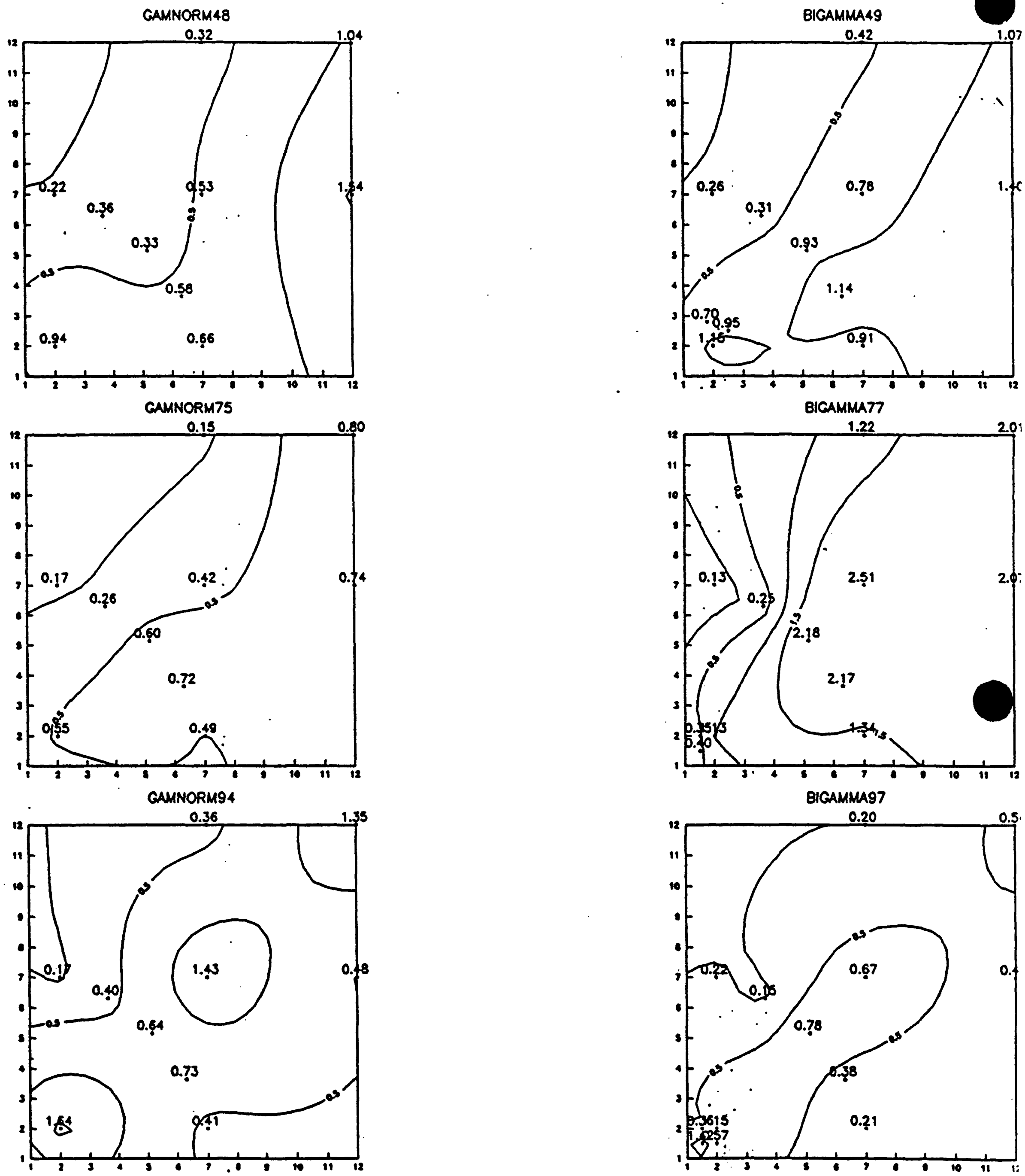


Figure 1. Comparison of Efficiency of Random-order, *vps* to Fixed-order(x) Sampling.

(Values plotted are $V(\hat{T}_y)$ for fixed-order(x) divided by $V(\hat{T}_y)$ for random-order, *vps* sampling.)

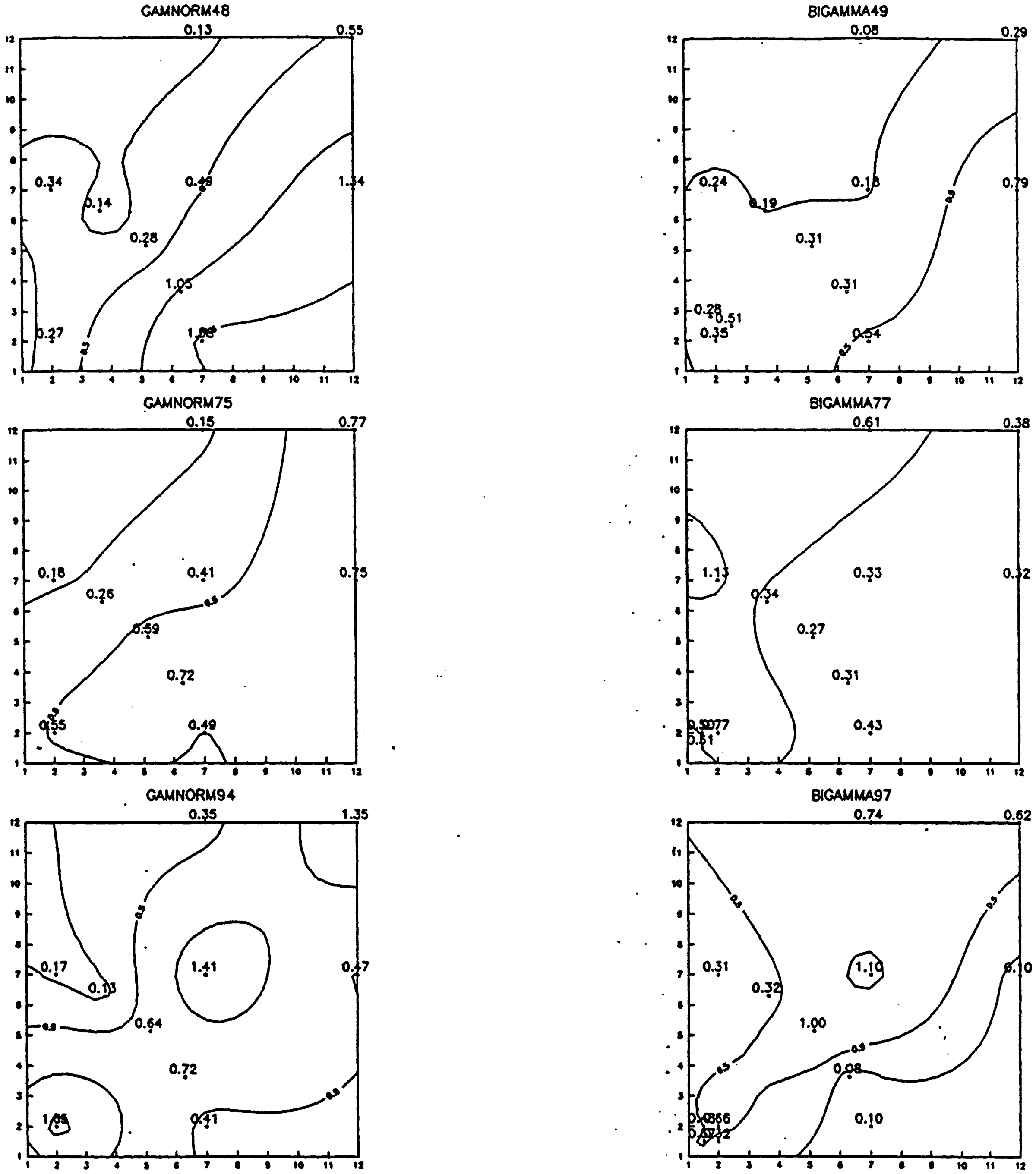


Figure 2. Comparison of Efficiency of Random-order, *vps* to Fixed-order(z) Sampling.

(Values plotted are $V(\hat{T}_y)$ for fixed-order(z) divided by $V(\hat{T}_y)$ for random-order, *vps* sampling.)

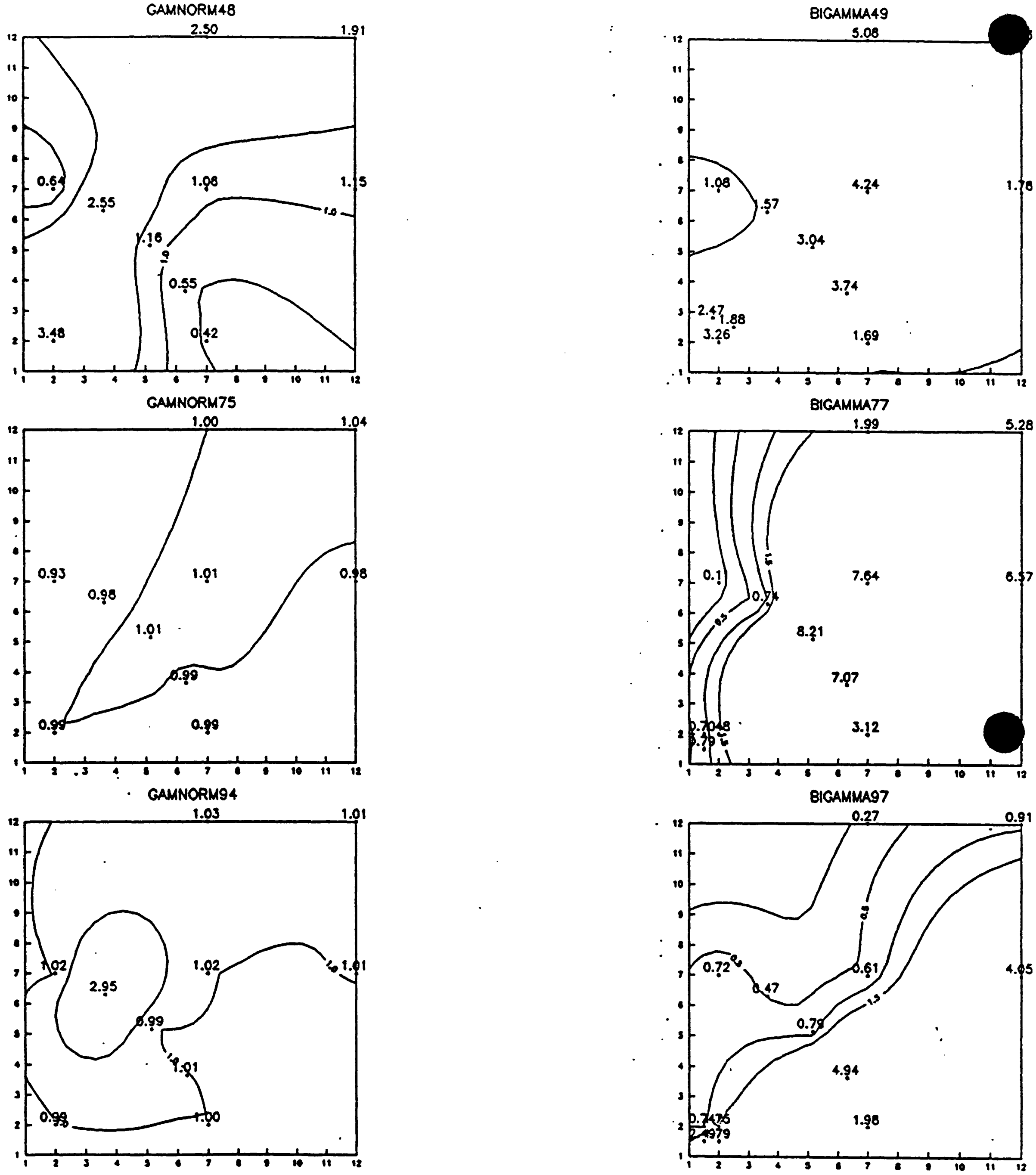


Figure 3. Comparison of Efficiency of Fixed-order(x) to Fixed-order(z) Sampling.

(Values plotted are $V(\hat{T}_y)$ for fixed-order(x) divided by $V(\hat{T}_y)$ for fixed-order(z).)

Figure 4. Relative Bias of Alternative Variance Estimators for Fixed-order(x) Sampling: BIGAMMA Family.

Column (1): $v^{o}_{\hat{Y}G}$ Column (2): $v^{*1}_{\hat{Y}G}$ Column (3): $v^{*2}_{\hat{Y}G}$

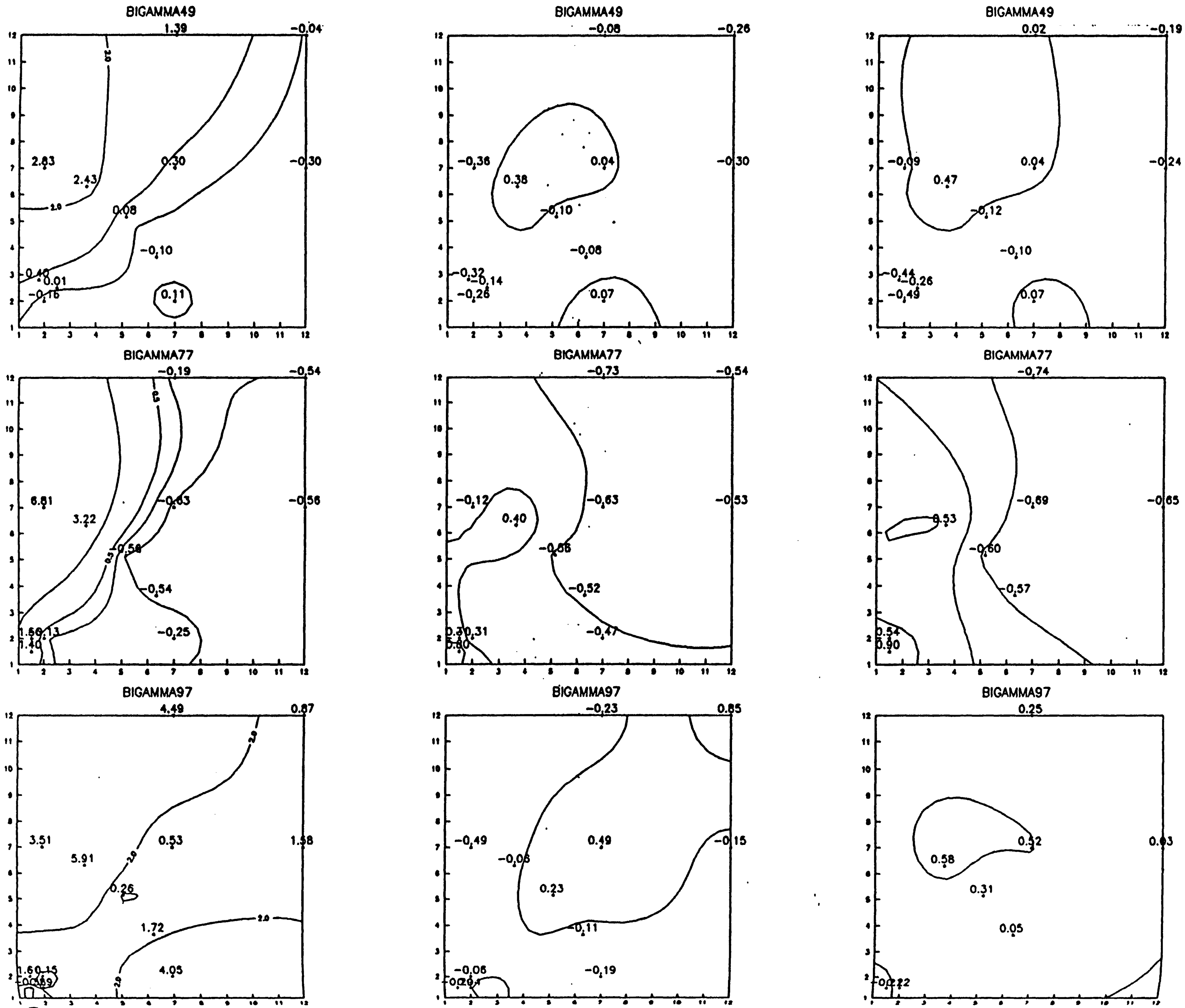
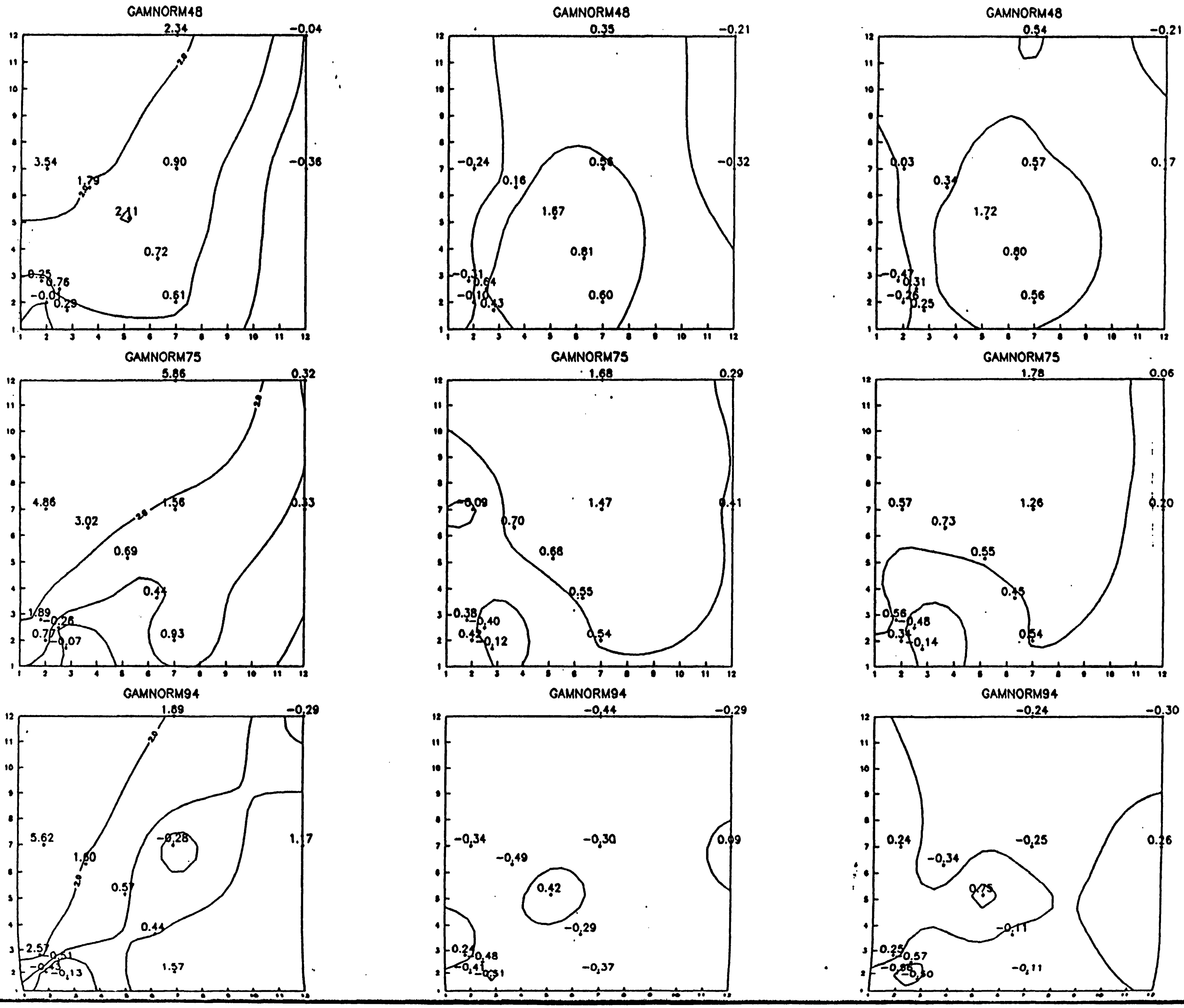

BIGAMMA49
1.39
-0.04
2.83
2.43
0.30
-0.30
0.08
-0.10
0.40
0.01
-0.16
0.11
BIGAMMA49
-0.08
-0.26
-0.36
0.38
0.04
-0.30
-0.10
-0.08
-0.32
-0.14
-0.26
0.07
BIGAMMA49
0.02
-0.19
-0.09
0.47
0.04
-0.24
-0.12
-0.10
-0.44
-0.26
-0.49
0.07
BIGAMMA77
-0.19
-0.54
6.81
3.22
-0.63
-0.56
-0.56
-0.54
-0.25
BIGAMMA77
-0.73
-0.54
-0.12
0.40
-0.63
-0.53
-0.52
-0.47
0.31
BIGAMMA77
-0.74
-0.69
-0.65
0.53
-0.60
-0.57
0.54
0.90
BIGAMMA97
4.49
0.87
3.51
5.91
0.53
1.58
0.26
1.72
4.05
BIGAMMA97
-0.23
0.85
-0.49
0.49
-0.15
0.23
-0.11
-0.06
-0.19
BIGAMMA97
0.25
0.58
0.52
0.03
0.31
0.05

Figure 5. Relative Bias of Alternate Variance Estimators for Fixed-order(z) Sampling: BIGAMMA Family.

Column (1): v^{o}_{YG} Column (2): v^{s1}_{YG} Column (3): v^{s2}_{YG}


GAMNORM48
2.34
-0.04
3.54
1.79
0.90
-0.36
2.11
0.72
0.25
0.76
-0.01
0.29
0.61
2.0
GAMNORM48
0.35
-0.21
-0.24
0.16
0.56
-0.32
1.67
0.81
-0.31
0.64
-0.10
0.43
0.60
GAMNORM48
0.54
-0.21
0.03
0.34
0.57
1.72
0.80
-0.47
0.31
-0.26
0.25
0.56
GAMNORM75
5.86
0.32
4.86
3.02
1.56
0.33
0.69
0.44
1.89
-0.26
0.70
-0.07
0.93
2.0
GAMNORM75
1.68
0.29
-0.09
0.70
1.47
0.41
0.68
0.55
0.38
0.40
0.42
0.12
0.54
GAMNORM75
1.78
0.06
0.57
0.73
1.26
0.20
0.55
0.45
0.56
0.48
0.34
-0.14
0.54
GAMNORM94
1.89
-0.29
5.62
1.60
-0.28
1.17
0.57
0.44
2.57
-0.51
-0.43
-0.13
1.57
2.0
GAMNORM94
-0.44
-0.29
-0.34
-0.49
-0.30
0.09
0.42
-0.29
-0.24
0.48
-0.41
-0.51
-0.37
GAMNORM94
-0.24
-0.30
0.24
-0.34
-0.25
0.26
0.75
-0.11
0.25
0.57
-0.86
-0.50
-0.11

9. REFERENCES

Bellhouse, D. R. (1988). Systematic Sampling. Handbook of Statistics, Vol. 6, P. R. Krishnaiah and C. R. Rao, eds., 125-145. Elsevier Science Publishers: Amsterdam.

Bourdeau, P. F. (1953). A test of random versus systematic ecological sampling. Ecology 34, 499-512.

Brewer, K. R. W., and Hanif, M. (1983). Sampling with Unequal Probabilities. Springer-Verlag: New York.

Buckland, W. R. (1951). A review of the literature of systematic sampling. J. Roy. Statist. Soc. Ser. B 13, 208-215.

Cochran, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of population. Ann. Math. Statist. 17, 164-177.

Cochran, W. G. (1977). Sampling Methods (3rd Edition). Wiley: New York.

Das, A. C. (1950). Two-dimensional systematic sampling and the associated stratified and random sampling. Sankhya 10, 95-108.

Freese, F. (1962). Elementary Forest Sampling. U. S. Department of Agriculture Handbook No. 232.

Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probability and without replacement. Ann. Math. Statist. 33, 350-374.

Hendricks, W. A. (1956). The Mathematical Theory of Sampling. The Scarecrow Press: New Brunswick, New Jersey.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47, 663-685.

Iachan, R. (1982). Systematic sampling: A critical review. Internat. Statist. Rev. 50, 293-303.

Isaki, C. T., and Pinciaro, S. J. (1977). Numerical comparison of some estimators of variance under PPS systematic sampling. Proceedings of the Survey Research Section, American Statistical Association, pp. 308-313.

Kuk, A. Y. C. (1989). Double bootstrap estimation of variance under systematic sampling with probability proportional to size. J. Statist. Comput. Simul. 31, 73-82.

Madow, W. G. (1949). On the theory of systematic sampling, II. Ann. Math. Statist. 20, 333-354.

Madow, W. G., and Madow, L. H. (1944). On the theory of systematic sampling, I. Ann. Math. Statist. 15, 1-24.

Milne, A. (1959). The centric systematic area-sample treated as a random sample. Biometrics 15, 270-297.

Murthy, M. N. (1967). Sampling Theory and Methods. Statistical Publishing Society: Calcutta.

von Neuman, J., Kent, R. H., Bellinson, H. R., and Hart, B. I. (1941). The mean square successive difference. Ann. Math. Statist. 12, 153-162.

Osborne, J. G. (1942). Sampling errors of systematic and random surveys of cover-type areas. J. Amer. Statist. Assoc. 37, 256-264.

Overton, W. S. (1964). On Estimating the Mean Ordinate of a Continuous Function over a Specified Interval. Ph.D. Thesis, North Carolina State University.

Overton, W. S. (1985). A Sampling Plan for Streams in the National Stream Survey. Technical Report 114, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Payandeh, B. (1970). Relative efficiency of two-dimensional systematic sampling. Forest Science 16, 271-276.

Poole, R. W. (1974). An Introduction to Quantitative Ecology. McGraw-Hill: New York.

Quenouille, M. H. (1949). Problems in plane sampling. Ann. Math. Statist. 20, 355-375.

Schreuder, H. T., Sedransk, J., and Ware, K. D. (1971). Forest sampling in the absence of a well-defined frame. In Statistical Ecology, Vol. 2. Sampling and Modelling Biological Populations and Population Dynamics, G. P. Patil, E. C. Pielou, and W. E. Waters (eds), International Cooperative Publishing House.

Stehman, S. V., and Overton, W. S. (1987a). An Empirical Investigation of the Variance Estimation Methodology Prescribed for the National Stream Survey: Simulated Sampling from Data Sets Obtained in the Pilot Study. Technical Report 118, Department of Statistics, Oregon State University, Corvallis, Oregon, 97331.

Stehman, S. V., and Overton, W. S. (1987b). A comparison of Variance Estimators of the Horvitz-Thompson Estimator in Random-Order, Variable Probability, Systematic Sampling. Biometrics Unit Manuscript Bu-935-M, Cornell University.

Stehman, S. V., and Overton, W. S. (1989). An Empirical, General Population Assessment of the Properties of Variance Estimators of the Horvitz-Thompson Estimator under Random-Order, Variable Probability Systematic Sampling. Biometrics Unit Manuscript Bu-936-M, Cornell University.

Wolter, K. M. (1985). Introduction to Variance Estimation. Springer-Verlag: New York.

Yates, F. (1949). Systematic sampling. Phil. Trans. Royal Soc. A 241, 345-377.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. J. Roy. Statist. Soc. Ser. B 15, 235-261.

Zinger, A. (1980). Variance estimation in partially systematic sampling. J. Amer. Statist. Assoc. 75, 206-211.